

# Parameter Estimation with BEAMS in the presence of biases and correlations

J. Newling<sup>1,2\*</sup>, B. Bassett<sup>1,2,3</sup>, R. Hlozek<sup>4</sup>, M. Kunz<sup>5</sup>, M. Smith<sup>2,6</sup>, M. Varughese<sup>7</sup>

<sup>1</sup>*Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch 7701, South Africa*

<sup>2</sup>*African Institute for Mathematical Sciences, 6-8 Melrose Road, Muizenberg, 7945, South Africa*

<sup>3</sup>*South African Astronomical Observatory, PO Box 9, Observatory 7935, South Africa*

<sup>4</sup>*Department of Astrophysics, Oxford University, Oxford OX1 3RH, United Kingdom*

<sup>5</sup>*Département de Physique Théorique, Université de Genève, Genève CH1211, Switzerland*

<sup>6</sup>*Astrophysics, Cosmology and Gravity Centre, University of Cape Town, Rondebosch 7701, South Africa*

<sup>7</sup>*Department of Statistical Sciences, University of Cape Town, Rondebosch 7701, South Africa*

Not submitted to arXiv at last compile: 31 October 2011

## ABSTRACT

The original formulation of BEAMS - Bayesian Estimation Applied to Multiple Species - showed how to use a dataset contaminated by points of multiple underlying types to perform unbiased parameter estimation. An example is cosmological parameter estimation from a photometric supernova sample contaminated by unknown Type Ibc and II supernovae. Where other methods require data cuts to increase purity, BEAMS uses all of the data points in conjunction with their probabilities of being each type. Here we extend the BEAMS formalism to allow for correlations between the data and the type probabilities of the objects as can occur in realistic cases. We show with simple simulations that this extension can be crucial, providing a 50% reduction in parameter estimation variance when such correlations do exist. We then go on to perform tests to quantify the importance of the type probabilities, one of which illustrates the effect of biasing the probabilities in various ways. Finally, a general presentation of the selection bias problem is given, and discussed in the context of future photometric supernova surveys and BEAMS, which lead to specific recommendations for future supernova surveys.

**Key words:** BEAMS, supernova, classification, typing, machine learning, selection bias, biased probabilities, Bayesian

## 1 INTRODUCTION

Type Ia Supernovae (SNeIa) provided the first widely accepted evidence for cosmic acceleration in the late 1990s (Riess et al. 1998; Perlmutter et al. 1999). While they were based on relatively small numbers of spectroscopically-confirmed SNeIa, those results have since been confirmed by independent analyses of other data sets (Eisenstein et al. 2005; Percival et al. 2007; Mantz et al. 2010; Fu et al. 2008; Giannantonio et al. 2008; Percival et al. 2010; Komatsu et al. 2011).

Next generation SN surveys such as LSST will be fundamentally different, yielding thousands of high-quality candidates every night for which spectroscopic confirmation will probably be impossible. Creating optimal ways of using this excellent photometric data is a key challenge in SN cosmology for the coming decade. There are two ways that one can imagine using photometric candidates. The first approach is to try to classify the candidates into Ia, Ibc or II SNe (Johnson & Crotts 2006; Kuznetsova & Connolly 2007; Poznanski et al. 2007; Rodney & Tonry 2009) and then use only those objects that are believed to be SNeIa above some threshold of confidence. This has recently been discussed by Sako et al. (2011) who showed that photometric cuts could achieve high purity. Nevertheless it is clear that this approach can still lead to biases and systematic errors from the small contaminating group when used in conjunction with the simplest parameter estimation approaches such as the maximum likelihood method.

A second approach is to use all the SNe, irrespective of how likely they are to actually be a SNIa. This is the approach exemplified by the BEAMS formalism, which accounts for the contamination from non-Ia SN data using the appropriate Bayesian framework, as presented in Kunz et al.

\* E-mail: james.newling@gmail.com

(2007), hereafter referred to as KBH. In KBH, two threads are woven: a general statistical framework, and a discussion of how it may be applied to SNeIa. As noted in KBH, the general framework can be applied to any parameter estimation problem involving several populations, and indeed may have already been done so in other fields. In this paper we take the same approach as in KBH of keeping the notation general enough for application to other problems, while discussing its relevance to SNe.

We will attempt to use the same notation as in KBH, but differ where we consider it necessary. For example, we write conditional probability functions as  $f_{\Theta|D}(\theta|d)$ . The quantity  $f_{\Theta|D}(\theta|d)\Delta\theta$ , should be interpreted as the probability that  $\Theta$  lies in the interval  $(\theta, \theta + \Delta\theta)$ , conditional on  $D = d$  (for small  $\Delta\theta$ ).

We preserve capital letters for random variables and lowercase letters for their observed values. In the BEAMS framework, one wishes to estimate parameter(s)  $\Theta$  from  $N$  observations of the random variable  $X$ . We will use the boldface  $\mathbf{X}$  to denote a vector of  $N$  such random variables:  $\mathbf{X} = X_{1\dots N}$ . An observation of  $X$  we will denote by  $x$ , so that the full set of  $N$  observations is denoted by  $\mathbf{x} = x_{1\dots N}$ . For SNe, the observations  $\mathbf{x}$  are the photometric data of the  $N$  SNe. As such, for SNe the probability density function (pdf)  $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$  is the likelihood of observing the photometric data  $\mathbf{x}$  assuming some cosmological parameters  $\theta$ , which we will discuss. The relationship between raw photometric data ( $X$ ) and the true cosmological parameters ( $\Theta$ ) is highly intricate, resulting in a pdf which cannot realistically be worked with, and so one first reduces each observation  $x$  to a single feature  $d$  for which there is a direct  $\Theta$ -dependent model. For SNe, if the parameters  $\Theta$  are for example  $\Omega_\Lambda$  and  $\Omega_m$ , then  $d$  will consist of an estimated luminosity distance and redshift. If the parameter of interest  $\Theta$  is a luminosity distance at a given redshift, then  $d$  will be simply a fitted distance modulus. Unless stated otherwise, this is the case.

The correct treatment of redshifts will be important to BEAMS as applied to future SN surveys. Future surveys will likely have only photometric information for the SNe but will have a spectroscopic redshift for the host galaxy obtained by chance (because of overlap with existing surveys) or through a targeted follow-up program. The SDSS-II supernova survey (Abazajian et al. 2009) is an example of both of these. There were host redshifts available from the main SDSS galaxy sample and there was also a targeted host followup program as part of the BOSS survey. Future large galaxy surveys like SKA, EUCLID or BigBOSS will likely provide a very large number of host galaxy redshifts for free.

BEAMS is unique in that the underlying types of the observations are not assumed known. In the case where there are two underlying types ( $T \in \{A, B\}$ ), each observation has an associated type probability ( $P$ ) of being type  $A$ ,

$$P \stackrel{\text{def}}{=} P(T = A|X_P),$$

where  $X_P$  is a subset of features of  $X$ . In other words,  $X_P$  is the component of the raw data  $X$  on which type probabilities are conditional. Note that we treat  $P$  as a random variable: while the value of  $P$  is completely determined by  $X_P$ , which in turn is completely determined by  $X$ ,  $X$  is a random variable and therefore so too is  $P$ . The realizations of the type probabilities  $\mathbf{P}$  of the  $N$  observations are denoted

by  $\mathbf{p} = p_{1\dots N}$ , and we will call them  $\tau_A$ -probabilities. The  $\tau_A$ -probability for a SN is thus the probability of being type Ia, conditional on knowing the subset  $x_P$  of the photometric data.  $x_P$  may be the full photometric time-series, the earliest segment of the SN's light curve, a fitted shape parameter, or any other extracted photometric information.

Finally, we mention that the type of the SN ( $T$ ) is a random variable with realisation denoted  $\tau$ . A summary of all the variables used in the paper is given in Table 1.

Attempts to approximate  $\tau_A$ -probabilities include those of Poznanski et al. (2002); Newling et al. (2011); Richards et al. (2011) and as implemented in SALT2 (Guy et al. 2007). Note that values obtained using these methods are only approximations of  $\tau_A$ -probabilities, as the algorithms are trained on only a handful of spectroscopically confirmed SNe. Note too that there is no sense in which one set of  $\tau_A$ -probabilities is *the* correct set, as this depends on what  $X_P$  is. Obtaining unbiased estimates of  $\tau_A$ -probabilities is not easy, and we will consider the problems faced in doing so in Section 7. For SNe, the problem is made especially difficult by the fact that spectroscopically confirmed SNe, which are used to train  $\tau_A$ -probability estimating algorithms, are brighter than unconfirmed photometric SNe.

In 2009 the Supernova Photometric Classification Challenge (SNPCC) was run to encourage work on SN classification by lightcurves alone (Kessler et al. 2010). Performance of the classification algorithms was judged according to the final purity and efficiency of extracted Ia samples. While the processing of photometric data is essential to the workings of BEAMS for SNe, the classification of objects is not required. It would be interesting to hold another competition where entrants are required to calculate  $\tau_A$ -probabilities for SNe. Algorithms would then not only need to recognise SNeIa, but would also need to provide precise, unbiased probabilities of the object being an SNeIa.

In brief, this paper consists of three more or less independent parts. In Section 2 we present an extension of BEAMS to the case where certain correlations, which were ignored in KBH, are present. In Section 3, we discuss the relevance of  $\tau_A$ -probabilities in a broader context, and specifically the importance to of them in BEAMS. Then in Sections 4, 5 and 6, we perform simulations to better understand the importance of sample sizes, nearness of population distributions, biases of  $\tau_A$ -probabilities and decisiveness of  $\tau_A$ -probabilities (to be defined). Finally, in Section 7 we present new ideas from the machine learning literature describing when and how  $\tau_A$ -probability biases emerge and how to correct for them. This is then discussed in the context of the SNPCC in Section 8.

## 2 INTRODUCING AND MODIFYING THE BEAMS EQUATIONS

The posterior probability on the parameter(s)  $\Theta$ , given the data  $\mathbf{D}$ , is derived in Section II of KBH as

$$f_{\Theta|D}(\theta|d) \propto f_{\Theta}(\theta) \times \sum_{\tau \in [A, B]^N} f_{D|\Theta, \tau}(d|\theta, \tau) \prod_{\tau_i=A} p_i \prod_{\tau_j=B} (1-p_j), \quad (1)$$

Random Variables		
R.V.	Data	Definition
$P$	$p$	The probability of being type $A$ conditional on $X_P$ . We call $P$ the $\tau_A$ -probability.
$D$	$d$	A particular feature of an object whose distribution depends directly on the parameter(s) we wish to approximate using BEAMS. SNe: $D$ is luminosity distance.
$T$	$\tau$	The type of an object, $T \in \{A, B\}$ SNe: $T \in \{\text{Ia}, \text{nIa}\}$
$X$	$x$	All the features observed of an object. SNe: $X$ is the photometric data.
$X_F$	$x_F$	That part of the features which affects confirmation probability. SNe: $X_F$ are peak apparent magnitudes.
$X_P$	$x_P$	That part of the features used to determine the $\tau_A$ -probability. SNe: $X_P$ could be any reduction of $X$ .
$F$	$f$	Whether the object is confirmed or not. For SNe: $F = 1$ if a spectroscopic confirmation is performed.
$\bar{P}$	$\bar{p}$	Is exactly $P$ if the object is unconfirmed and 1 or 0 if confirmed, depending on type.

**Table 1.** A description of all the random variables used in this paper.

where the  $p_i$ s are  $\tau_A$ -probabilities. The summation is over all of the  $2^N$  possible ways that the  $N$  observations can be classified into two classes. We will refer to the expression on the right of (1) as the KBH posterior. When the  $N$  observations are assumed to be independent, that is when

$$f_{D|\Theta, T}(\mathbf{d}|\theta, \tau) = \prod_{i=1}^N f_{D_i|\Theta, T_i}(d_i|\theta, \tau_i),$$

the KBH posterior reduces,

$$\prod_{i=1}^N [f_{D_i|\Theta, T_i}(d_i|\theta, A) p_i + f_{D_i|\Theta, T_i}(d_i|\theta, B) (1 - p_i)]. \quad (2)$$

There is one substitution in the derivation of the KBH posterior on which we would like to focus, given in KBH as eqn. (5) on page 3:

$$f_T(\tau) = \prod_{\tau_i=A} p_i \prod_{\tau_i=B} (1 - p_i). \quad (3)$$

Equation (3) states that the l.h.s. prior probability of the SNe having types  $\tau$  is given by the product on the r.h.s. involving  $\tau_A$ -probabilities. We argue that this product should not be treated as the prior  $f_T$ , but rather as the conditional  $f_{T|P}$ . In effect, we argue that KBH should not use the  $\tau_A$ -probabilities  $p$  unless  $P$  is explicitly included as a conditional parameter. It is to this end that we now re-derive the posterior on  $\Theta$ , taking  $f_{\Theta|D, P}(\theta|\mathbf{d}, \mathbf{p})$  as a starting point, discussing at each line what has been used.

$$f_{\Theta|D, P}(\theta|\mathbf{d}, \mathbf{p})$$

→ We will first use the definition of conditional probability to obtain,

$$= \frac{f_{\Theta, D, P}(\theta, \mathbf{d}, \mathbf{p})}{f_{D, P}(\mathbf{d}, \mathbf{p})}.$$

→ The term in the numerator can then be written as the sum over all  $2^N$  possible type vectors,

$$= \sum_{\tau} \frac{f_{\Theta, D, P, T}(\theta, \mathbf{d}, \mathbf{p}, \tau)}{f_{D, P}(\mathbf{d}, \mathbf{p})}.$$

→ The numerator is again modified using the definition of conditional probability,

$$= \sum_{\tau} \frac{f_{D|\Theta, P, T}(\mathbf{d}|\theta, \mathbf{p}, \tau) f_{\Theta, P, T}(\theta, \mathbf{p}, \tau)}{f_{D, P}(\mathbf{d}, \mathbf{p})}.$$

→ We will now assume that the probability of having  $\tau_A$ -probabilities and types  $\mathbf{p}$  and  $\tau$  respectively are independent of  $\Theta$ . As noted following eqn.(4) in KBH, for SNe this assumption rests on the fact that  $\Theta$  (that is  $\Omega_m, \Omega_\Lambda$ ) describes large scale evolution, while the SN types  $\tau$  depend on local astrophysics, with little or no dependence on perturbations in dark matter.

$$= \sum_{\tau} \frac{f_{D|\Theta, P, T}(\mathbf{d}|\theta, \mathbf{p}, \tau) f_{\Theta}(\theta) f_{P, T}(\mathbf{p}, \tau)}{f_{D, P}(\mathbf{d}, \mathbf{p})}.$$

→ Rearranging this, and again using the definition of conditional probability, we obtain,

$$= \frac{f_P(\mathbf{p})}{f_{D, P}(\mathbf{d}, \mathbf{p})} f_{\Theta}(\theta) \sum_{\tau} f_{D|\Theta, P, T}(\mathbf{d}|\theta, \mathbf{p}, \tau) f_{T|P}(\tau|\mathbf{p}).$$

→ The first term on the line above is constant with respect to  $\Theta$ , and so is absorbed into a proportionality constant. We now make one final weak assumption:  $f_{T|P}(\tau|\mathbf{p}) = \prod_{i=1}^N f_{T_i|P_i}(\tau_i|p_i)$ . This assumption will be necessary to make a comparison with the KBH posterior. Making this assumption we arrive at,

$$\propto f_{\Theta}(\theta) \sum_{\tau} f_{D|\Theta, P, T}(\mathbf{d}|\theta, \mathbf{p}, \tau) \prod_{\tau_i=A} p_i \prod_{\tau_j=B} (1 - p_j). \quad (4)$$

We will refer to the newly derived expression (4) as the full posterior. Let us now consider the difference between the KBH posterior (1) and the full posterior, and notice that in the full posterior, the likelihood of the data  $\mathbf{D}$  is conditional on  $\Theta, \mathbf{P}$  and  $\mathbf{T}$ , whereas in the KBH posterior  $\mathbf{D}$  is only conditional on  $\Theta$  and  $\mathbf{T}$ . This is the only difference between the two posteriors, and so when  $\mathbf{D}|\Theta, \mathbf{T}$  is independent of  $\mathbf{P}$ , the posterior (4) reduces to the KBH posterior (1), making them equivalent. This is an important result: when  $\mathbf{D}|\Theta, \mathbf{T}$  and  $\mathbf{P}$  are independent, the KBH and full posteriors are the same.

Our results can be summarised as follows,

(1) As the posterior  $f_{\Theta|D}(\theta|\mathbf{d})$  is not conditional on  $\tau_A$ -probabilities it should be independent of  $\tau_A$ -probabilities, and we thus prefer to replace the KBH posterior in (1) by

$$f_{\Theta|D}(\theta|\mathbf{d}) \propto f_{\Theta}(\theta) \times \sum_{\tau \in [A, B]^N} f_{D|\Theta, T}(\mathbf{d}|\theta, \tau) \prod_{\tau_i=A} \pi \prod_{\tau_j=B} (1 - \pi),$$

where  $\pi$  is an estimate of the global proportion of type  $A$  objects.

(2)  $f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p})$  is always given by the full posterior (4). When  $\mathbf{D}|\Theta, \mathbf{T}$  and  $\mathbf{P}$  are independent, it reduces to the KBH posterior (1).

It is worth discussing for SNe the statement, “ $\mathbf{D}|\Theta, \mathbf{T}$  and  $\mathbf{P}$  are not independent”. One incorrect interpretation of this statement is, “given that we know the cosmology is  $\Theta$ , observing<sup>1</sup>  $\mathbf{P}$  for a SN of unknown type adds no information to the estimation of the distance modulus.” Indeed it is difficult to imagine how this could be the case: we know that SNIa are brighter than other SNe, and therefore obtaining a  $\tau_A$ -probability close to 1 shifts the estimated distance modulus downwards (towards being brighter).

A correct interpretation of the statement is, “given the cosmology  $\Theta$ , observing  $\mathbf{P}$  of a SN of known type adds no information to the estimation of the distance modulus.” It may seem necessarily true that a  $\tau_A$ -probability contributes no new information if the type of the SN is already known, but this is not in general the case; it depends on the method by which  $\tau_A$ -probabilities are obtained.

Currently for SNe, fitted distance moduli and approximations of  $\tau_A$ -probabilities are frequently obtained simultaneously, using for example SALT2 (Guy et al. 2007). This in itself suggests that  $\mathbf{D}|\Theta, \mathbf{T}$  and  $\mathbf{P}$  will not be independent. In some cases however,  $\tau_A$ -probabilities are calculated from the early stages of the light curves (Sullivan et al. 2006; Sako et al. 2008) while the distance modulus is estimated from the peak of the light curve, and so the dependence may be weak. As another example, in Section 4.4 of Newling et al. (2011)  $\tau_A$ -probabilities are obtained directly from a Hubble diagram. Objects lying in regions of high relative SNIa density are given higher  $\tau_A$ -probabilities than objects lying in low relative SNIa density. As a result, at a given redshift, brighter nIa SNe have higher  $\tau_A$ -probabilities than faint nIa SNe. Similarly, at a given fitted distance modulus (fitted assuming type Ia), nIa will lie on average at lower redshifts than Ia. Both of these cases, (distance modulus |  $\Theta$ , type) being correlated with  $P$ , and (redshift |  $\Theta$ , type) being correlated with  $P$ , are precisely when  $\mathbf{D}|\Theta, \mathbf{T}$  and  $\mathbf{P}$  are dependent. In Section 6 a simulation illustrating this dependence is presented.

For completeness, we mention that in the case of independent observations, that is when,

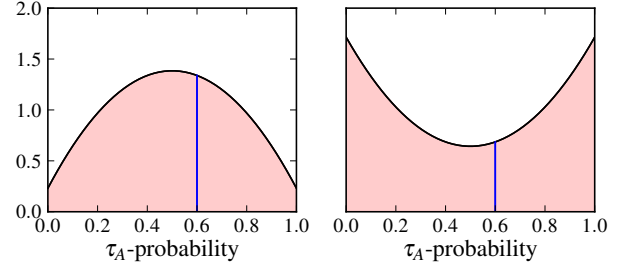
$$f_{\mathbf{D}|\Theta,\mathbf{P},\mathbf{T}}(\mathbf{d}|\theta,\mathbf{p},\boldsymbol{\tau}) = \prod_{i=1}^N f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,\tau_i),$$

the full posterior (4) reduces to,

$$f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p}) \propto \prod_{i=1}^N [f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,A)p_i + f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,B)(1-p_i)]. \quad (5)$$

In Section 7 we will make suggestions as to what

<sup>1</sup> Of course we mean “observing” in the statistical sense, that is obtaining the realisation of the  $\tau_A$ -probability ( $p$ ) with some software



**Figure 1.** Two  $\tau_A$ -probability distributions, both with means of 0.5. Using a threshold of 0.6, we have on left: FPR = 0.17, FNR = 0.45 and on right: FPR: 0.15, FNR = 0.28

functional form may be chosen for  $f_{D_i|\Theta,P_i,T_i}$  when using BEAMS for independent SNe.

### 3 RATING $\tau_A$ -probabilities

An object’s  $\tau_A$ -probability is the expected proportion of other objects with its features which are type  $A$ . In other words, if an object has features  $x$ , its  $\tau_A$ -probability is the expected proportion of objects with features  $x$  which are type  $A$ . Suppose that the global distribution of  $P$  is  $f_P$ . The expected total proportion of type  $A$  objects is then

$$P(T = A) = \langle P \rangle = \int_0^1 p f_P(p) dp. \quad (6)$$

In some circumstances, it is necessary to go beyond calculating  $\tau_A$ -probabilities and commit to an absolute classification, as was the case in the SNPCC. In such cases the optimal strategy moving from a  $\tau_A$ -probability to an absolute type ( $A$  or  $B$ ) is to classify objects positively ( $A$ ) when the  $\tau_A$ -probability is above some threshold probability ( $c$ ). The False Positive Rate (FPR) using such a strategy is

$$\begin{aligned} \text{FPR}(f_P) &= P(P > c | T = B) \\ &= \frac{\int_c^1 (1-p) f_P(p) dp}{\int_0^1 (1-p) f_P(p) dp}, \end{aligned} \quad (7)$$

and the False Negative Rate is

$$\begin{aligned} \text{FNR}(f_P) &= P(P < c | T = A) \\ &= \frac{\int_0^c p f_P(p) dp}{\int_0^1 p f_P(p) dp}. \end{aligned} \quad (8)$$

For SNe the FPR is the proportion of nIa SNe which are misclassified, while the FNR is the proportion of SNIa which are misclassified (missed).

Intuition dictates that for classification problems, a useful  $f_P$  will be one whose mass predominates around 0 and 1. That is, an  $f_P$  which with high probability attaches decisive<sup>2</sup>  $\tau_A$ -probabilities to observations. To minimize the FPR and FNR this is optimal, as illustrated in Figure 1.

We will be presenting a simulation illustrating how the decisiveness of  $\tau_A$ -probabilities affects the parameter estimation of BEAMS. To simplify our study of the effect of the decisiveness of  $\tau_A$ -probabilities on BEAMS, we introduce a

<sup>2</sup> we say  $p_1$  is more decisive than  $p_2$  if  $|p_1 - 0.5| > |p_2 - 0.5|$ .

family of distributions: For each  $\mathcal{P} \in [0.5, 1]$  we have the distribution

$$f^{\mathcal{P}}(p) = \frac{1}{2} (\delta_{\mathcal{P}}(p) + \delta_{1-\mathcal{P}}(p)) \quad (9)$$

where  $\delta_{\mathcal{P}}$  and  $\delta_{1-\mathcal{P}}$  are  $\delta$ -functions centered at  $\mathcal{P}$  and  $1-\mathcal{P}$  respectively. It is worth mentioning that we will be drawing probabilities from this distribution, which is potentially confusing. Drawing a observation of  $P$  from (9) is equivalent to drawing it from  $\{1-\mathcal{P}, \mathcal{P}\}$  with equal probability:

$$P(P=p) = \begin{cases} 0.5 & \text{if } p = \mathcal{P} \\ 0.5 & \text{if } p = 1-\mathcal{P}. \end{cases}$$

If  $\mathcal{P}_1$  is more decisive than  $\mathcal{P}_2$ , we say that the distribution  $f^{\mathcal{P}_1}$  is more decisive than  $f^{\mathcal{P}_2}$ .

On page 5 of KBH it is stated that the expected proportion of type  $A$  objects (6) determines the expected error in estimating a parameter which is independent of population  $B$ . Specifically, they present the result that the expected error when estimating a parameter  $\mu$  with  $N$  objects using BEAMS is given by,

$$\sigma_{\mu} \propto \sqrt{\langle P \rangle N}. \quad (10)$$

It should be noted that the the result from KBH (10) is an asymptotic result in  $N$ . For small  $N$ , the decisiveness of the probabilities plays an important part. If (6) were the only factor determining the expected error ( $\sigma_{\mu}$ ), then  $f^{0.5}$  would be equivalent to  $f^1$  in terms of expected error. This would mean that perfect type knowledge does not reduce error, which would be surprising. An example in Section 4.1 illustrates that decisiveness does play a role in determining the error.

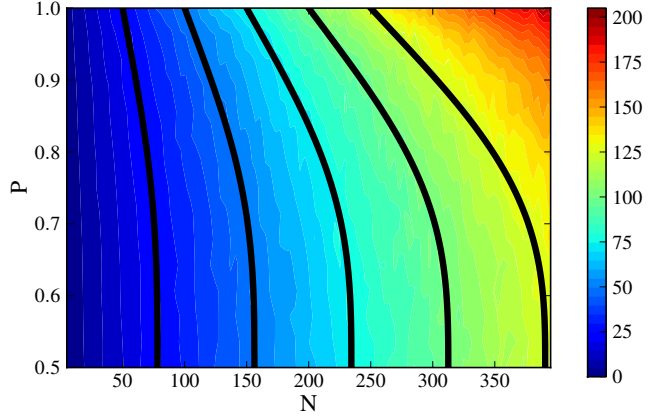
As mentioned on page 8 of KBH, the effect of biases in  $\tau_A$ -probabilities on BEAMS can be catastrophic. Therein they consider the case where there is a uniform bias ( $a$ ) of the  $\tau_A$ -probabilities. That is, if observation  $i$  has a claimed  $\tau_A$ -probability  $p_i$  of being type  $A$ , then there is a real probability  $p_i - a$  that it is type  $A$ . KBH show how, by including a free global shift parameter, such a bias is completely removed. However it is not clear what to do if the form of the bias is unknown. For example, it could be that there is an ‘over-confidence’ bias, where to obtain the true  $\tau_A$ -probabilities one needs to transform the claimed priors ( $\tilde{p}$ ) by

$$p = 0.2 + 0.6\tilde{p}. \quad (11)$$

Introducing a bias such as the one defined by (11) will have no effect on the optimal FPR and FNR, provided the probability threshold is chosen optimally. This is because (11) is a one-to-one biasing, and so a threshold ( $\tilde{c}$ ) on biased probabilities results in exactly the same partitioning as a threshold in the unbiased space of  $0.2 + 0.6\tilde{c}$ . However, introducing a bias such as (11) does have an effect on BEAMS parameter estimation, as we show in Section 5. In Section 7 we discuss how to guarantee that the  $\tau_A$ -probabilities are free of bias.

#### 4 EFFECTS OF DECISIVENESS AND SAMPLE SIZE ON BEAMS

In this section we will perform simulations to better understand the key factors in BEAMS. The data generated will



**Figure 2.** Contour plot of  $h(N, \mathcal{P})$ . The solid lines are approximations to lines of constant  $h$ , of the form (13).

have the following cosmological analogy:  $\Theta$  - distance modulus at a given redshift  $z_0$ ;  $\mathbf{d}$  - the fitted distance moduli of SNe at  $z_0$ . Furthermore,  $\mathbf{D}|\Theta, \mathbf{T}$  and  $\mathbf{P}$  will be independent, such that the KBH and full posterior are equivalent.

##### 4.1 Simulation 1: Estimating a population mean

This simulation was performed to see how the performance of BEAMS is affected by the decisiveness of  $\tau_A$ -probabilities, and by the size of the data set. The two populations ( $A$  and  $B$ ) were chosen to have distributions,

$$f_{D|T}(d, \tau) = \text{Normal}(\mu_{\tau}, 1), \quad (12)$$

where  $\mu_A = -1$  and  $\mu_B = +1$ , as illustrated in Figure 3. The  $\tau_A$ -probability distribution is chosen to be  $f^{\mathcal{P}}$ , so that about half of the observations have a  $\tau_A$ -probability of  $\mathcal{P}$ , with the remaining observations having  $\tau_A$ -probabilities of  $1-\mathcal{P}$ . By varying  $\mathcal{P}$  we vary the decisiveness.

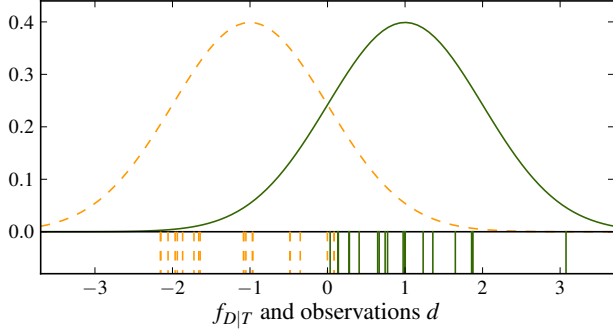
Let us make it clear how the data for this simulation is generated. First, a  $\tau_A$ -probability ( $p$ ) is selected to be either  $\mathcal{P}$  with probability 0.5 or  $1-\mathcal{P}$  with probability 0.5, that is according to  $f^{\mathcal{P}}$ . Second, the type of the observation is chosen, with probability  $p$  it is chosen as  $A$ , and with probability  $1-p$  it is chosen as  $B$ . Finally, the data ( $d$ ) is drawn from (12). Notice that  $D|T$  is independent of  $P$ , and so the KBH posterior is equivalent to the full posterior.

In this simulation we only estimate  $\mu_A$ , with all other parameters known. We use the following Figure of Merit to compare the performance with different sample sizes ( $N$ ) and decisivenesses ( $\mathcal{P}$ ):

$$h(N, \mathcal{P}) = \frac{1}{\langle \hat{\mu}_A - \mu_A \rangle^2},$$

where  $\hat{\mu}_A$  is the maximum likelihood estimate of  $\mu_A$  using the KBH posterior on a sample of size  $N$  with  $\tau_A$ -probabilities from  $f^{\mathcal{P}}$ , and  $\langle \cdot \rangle$  denotes an expectation. Values of  $h$  were obtained by simulation, illustrating in Figure 2 the performance of BEAMS for various  $(N, \mathcal{P})$  combinations. A good approximation to the FoM  $h$  in Figure 2 appears to be

$$h(N, \mathcal{P}) \approx N \left( 0.32 + 1.44 \left( \mathcal{P} - \frac{1}{2} \right)^2 \right), \quad (13)$$



**Figure 3.** Above are the population  $A$  (left) and population  $B$  (right) distributions, with (for Simulation 4.2) the observed values of  $D$  drawn from these distributions shown as vertical lines beneath.

although this is an ad hoc observation. One interesting observation is that  $h(N, \mathcal{P} = 1) \approx h(1.5N, \mathcal{P} = 0.5)$  in the region illustrated in Figure 2. This says that given a completely blind sample ( $\mathcal{P} = 0.5$ ), and the option to either double its size ( $N \rightarrow 2N$ ) or to discover the hidden types ( $\mathcal{P} : 0.5 \rightarrow 1$ ), doubling its size will provide more information about  $\mu_A$ . It is important to reiterate that, according to previously mentioned result of KBH, in the limit of  $N \rightarrow \infty$  we do not expect  $\mathcal{P}$  to play any part in determining  $h(N, \mathcal{P})$ . That is, for  $N$  sufficiently large, the FoM will be independent of  $\mathcal{P}$ .

While this simulation is too simple to make extrapolations about cosmological parameter estimations from, it may suggest that the information contained in unconfirmed photometric data may be currently underestimated.

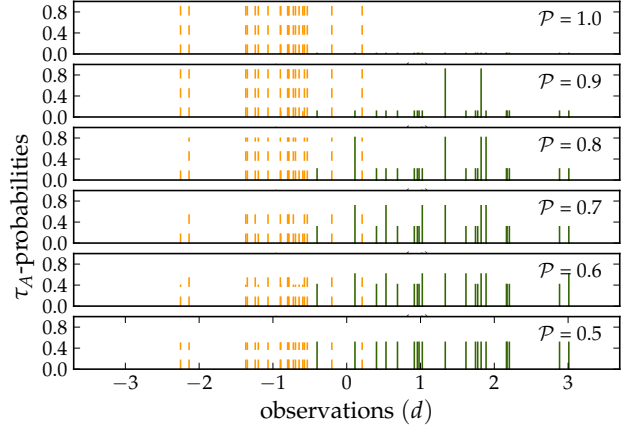
#### 4.2 Simulation 2: Estimating two population means

The two population distributions for this simulation are the same as those presented in Simulation 1 and as illustrated in Figure 3. In this simulation, we leave both the population means as free parameters to be fitted for. Twenty objects are drawn from the types  $A$  and  $B$ , with the  $\tau_A$ -probabilities are drawn from  $f^{\mathcal{P}}$ . The simulation is done with five different  $\mathcal{P}$  values. The  $\tau_A$ -probabilities are illustrated in Figure 4, and the approximate shape of the posterior marginals of  $\mu_A$  for each  $\mathcal{P}$  value are illustrated in Figure 5 by MCMC chain counts.

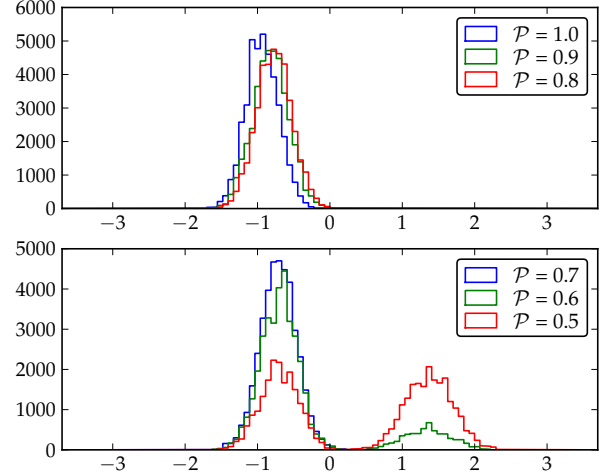
There are two interesting results from this simulation. The first is that there is negligible difference in performance between  $\mathcal{P} = 1$  and  $\mathcal{P} = 0.7$ , so that having a 30% type uncertainty for all objects as opposed to absolute type knowledge does not weaken the results. The second is that as  $\mathcal{P}$  approaches 0.5, BEAMS still correctly locates the population means but is unsure which mean belong to which population.

### 5 EFFECTS OF $\tau_A$ -PROBABILITIES BIAS ON BEAMS

In the previous section we considered the effect of the decisiveness of  $\tau_A$ -probabilities on the performance of BEAMS.



**Figure 4.** For values of  $\mathcal{P}$  from 1 (above) to 0.5 (below), a  $\tau_A$ -probability of  $\mathcal{P}$  or  $1 - \mathcal{P}$  is attached to each observation.



**Figure 5.** MCMC chain counts, approximating the posterior distributions of  $\mu_A$  for the different values of decisiveness,  $\mathcal{P}$ .

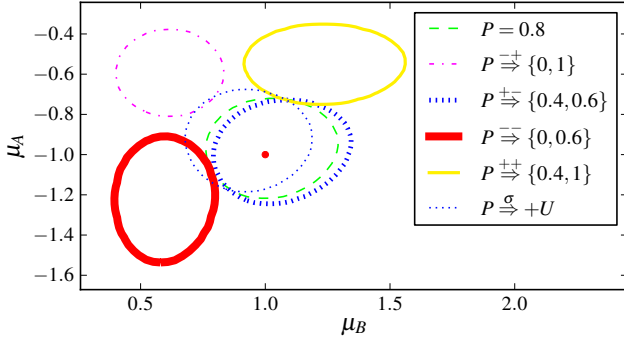
In this section we will consider the effect of using incorrect  $\tau_A$ -probabilities. We will again be estimating  $\mu_A$  and  $\mu_B$  where they are  $-1$  and  $1$  respectively, and the population variances are again both known to be 1. The true  $\tau_A$ -probability distribution will be  $f^{0.8}$ , that is

$$P(P = p) = \begin{cases} 0.5 & \text{if } p = 0.8 \\ 0.5 & \text{if } p = 0.2 \end{cases}$$

It is worth reminding the reader that we are drawing probabilities from a probability distribution, an unusual thing to do. To generate a  $\tau_A$ -probability from this distribution, one could flip a coin, and return  $p = 0.2$  if  $H$  and  $p = 0.8$  if  $T$ . We consider the effect of biasing  $\tau_A$ -probabilities generated in such a manner in the following ways:

- (i)  $\mathcal{P} \rightrightarrows^+ \{0, 1\}$ . Here the decisiveness of the  $\tau_A$ -probabilities





**Figure 6.** The 99 % posterior confidence regions using the five biasings of the  $\tau_A$ -probabilities, as described in Section 5.

is overestimated, so that  $p = 0.8 \rightarrow p = 1$  and  $p = 0.2 \rightarrow p = 0$ .

- (ii)  $\mathcal{P} \stackrel{\pm}{\Rightarrow} \{0.4, 0.6\}$ . Here the decisiveness of the  $\tau_A$ -probabilities is underestimated, so that  $p = 0.8 \rightarrow p = 0.6$  and  $p = 0.2 \rightarrow p = 0.4$ .
- (iii)  $\mathcal{P} \stackrel{\pm}{\Rightarrow} \{0, 0.6\}$ . Here the  $\tau_A$ -probabilities are underestimated by 0.2, so that  $p = 0.8 \rightarrow p = 0.6$  and  $p = 0.2 \rightarrow p = 0$ .
- (iv)  $\mathcal{P} \stackrel{\pm}{\Rightarrow} \{0.4, 1\}$ . Here the  $\tau_A$ -probabilities are overestimated by 0.2, so that  $p = 0.8 \rightarrow p = 1$  and  $p = 0.2 \rightarrow p = 0.4$ .
- (v)  $\mathcal{P} \stackrel{\pm}{\Rightarrow} U$ . Here, to each  $\tau_A$ -probability a uniform random number from  $[-0.2, 0.2]$  is independently added.

The 99% posterior confidence regions obtained using these biased  $\tau_A$ -probabilities in a simulation of 400 points are illustrated in Figure (6). The underestimation of decisiveness (ii) has little effect on the final confidence region, but overestimating the  $\tau_A$ -probability decisiveness (i) results in a  $6\sigma$  bias. Note that overestimating decisiveness results in the estimate  $(\hat{\mu}_A, \hat{\mu}_B)$  being biased towards  $(\mu_B, \mu_A)$ . This is caused by type  $B$  objects which are too confidently believed to be type  $A$ , which pull  $\hat{\mu}_A$  towards  $\mu_B$ , and type  $A$  objects which are too confidently believed to be type  $B$ , which pull  $\hat{\mu}_B$  towards  $\mu_A$ .

The contrast in effect between underestimating and overestimating the decisiveness of  $\tau_A$ -probabilities is interesting, and not easy to explain. One suggestion we have received is to consider the cause of the observed effect as being analogous to the increased contamination rate induced by overestimating the decisiveness in the case BEAMS is not used. With an increased contamination rate comes an increased bias, precisely as observed in Figure 6. It is worth mentioning that underestimating the decisiveness is not entirely without effect, as simulations with more pronounced drops in  $\mathcal{P}$  ( $0.95 \rightarrow 0.55$ ) result in noticeable increases in the size of the 99% confidence region.

The effect of the flat  $\tau_A$ -probability shifts (iii) and (iv) introduce biases larger than  $4\sigma$ . This case was considered in KBH where, as we have already mentioned, it was shown that simultaneously fitting for this bias completely compensates for it. While this is a pleasing result, one would prefer to know that the  $\tau_A$ -probabilities are correct, as one cannot be sure what form the biasing will take.

One phenomenon which is observed in this simulation, as it was in simulations as summarised in Table II on page

8 of KBH, is that a flat  $\tau_A$ -probability shift in confidence towards being type  $B$  (iii) does not bias the estimate of  $\mu_A$  as much as it does the estimate of  $\mu_B$ , and vica versa. In other words, underestimating the probabilities that objects are type  $A$  will result in less biased population  $A$  parameters than overestimating the probabilities. This result may also be understood in light of an analogy to increased contamination versus reduced population size in the case where BEAMS is not used.

Finally, we notice that in this simulation the addition of unbiased noise to the  $\tau_A$ -probabilities (v) has an insignificant effect. This suggests that systematic biases should be the primary concern of future work on the estimation of  $\tau_A$ -probabilities.

## 6 WHEN GIVEN TYPE, THE DATA IS STILL DEPENDENT ON $\tau$ -PRIORS

In this section we consider for the first time a simulation in which the data is not drawn from  $f_{D|T}$ , but from  $f_{D|T,P}$ , so that there is a dependence of the data on the  $\tau_A$ -probability even when the type is known. The conditional pdfs are shown in Figure 7. To clarify the difference between this simulation and the previous ones, prior to this data was simulated as follows:

$$P \rightarrow T|P \rightarrow D|T,$$

where at the last step, the data was generated with a dependence only on type. Now it will be simulated as:

$$P \rightarrow T|P \rightarrow D|P, T.$$

More specifically, to generate data we start by drawing a  $\tau_A$ -probability from  $f^{0.7}$ ,

$$P(P = p) = \begin{cases} 0.5 & \text{if } p = 0.7 \\ 0.5 & \text{if } p = 0.3. \end{cases}$$

Note that the above distribution guarantees that  $P(T = A) = \frac{1}{2}$ . When the  $\tau_A$ -probability ( $p$ ) has been generated, we draw a type ( $\tau$ ) from  $\{A, B\}$  according to

$$P(T = \tau) = \begin{cases} p & \text{if } \tau = A \\ 1 - p & \text{if } \tau = B. \end{cases}$$

Once we have  $p$  and  $\tau$ , we generate  $d$ . The marginals  $f_{D|P,T}(d|p, \tau)$  have been chosen such that we have

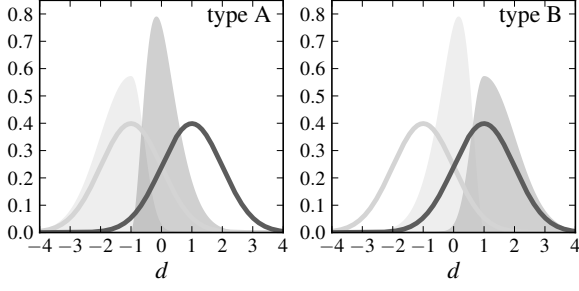
$$f_{D|T}(d|A) = \text{Normal}(-1, 1) \quad (14)$$

$$f_{D|T}(d|B) = \text{Normal}(1, 1), \quad (15)$$

as before. The marginal  $f_{D|P,T}(d|0.7, A)$  is composed of the halves of two Gaussian curves with different  $\sigma$ s, chosen such that the tail away from the  $B$  population is longer than the one towards the  $B$  population. Specifically,

$$\begin{aligned} f_{D|P,T}(d|0.7, A) &= \\ &= \begin{cases} K \exp -\frac{1}{2}(d+1)^2 & \text{if } d < -1 \\ K \exp -\frac{100}{32}(d+1)^2 & \text{if } d > -1 \end{cases} \end{aligned}$$

where  $K$  is a normalizing constant. The marginal  $f_{D|P,T}(d|0.3, A)$  is then constructed to guarantee (14). The above construction guarantees that the population of  $A$  objects with low  $\tau_A$ -probabilities (0.3) lie on average closer to the  $B$  mean than do objects with high (0.7)  $\tau_A$ -probabilities.



**Figure 7.** Plots of  $f_{D|P,T}(d|p,\tau)$  (filled curves) for  $p = 0.7$  (light) and  $p = 0.3$  (dark), and for type A (left) and type B (right). Overlying are  $f_{D|T}(d|A)$  (light) and  $f_{D|T}(d|B)$  (dark).

The marginals of the  $B$  population are constructed to mirror exactly the  $A$  population marginals, as illustrated in Figure 7.

To compare the use of the KBH BEAMS posterior (2) with the full conditional posterior (5), we randomly draw 40 data points from the above distribution and construct the respective posterior distributions, as illustrated in Figure 8. Observe that the KBH posterior is significantly wider than the full posterior. Indeed, approximately half of the interior of the 80% region of the KBH posterior is ruled out to 1% by the full posterior. It is interesting to note that, while the KBH posterior is wider than the full posterior, it is not biased. This result goes against our intuition; we believed that the KBH posterior would result in estimates for  $\mu_A$  and  $\mu_B$  which exaggerated  $|\mu_A - \mu_B|$ . Whether it is a general result that no bias exists when the KBH posterior is used, or if there can exist dependencies between  $P$  and  $D$  for which the use of (1) leads to a bias, remains an open question.

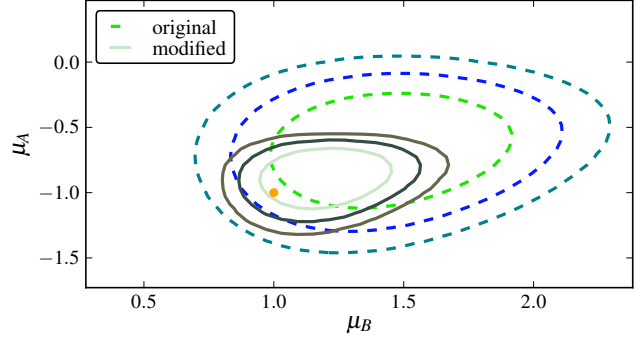
Figure 8 illustrates one realisation from the distribution we have described, but repeated realisations show that on average, the variance in the maximum likelihood estimator using the KBH posterior is  $\sim 3$  times larger than the variance using the modified posterior. While these simulations are too simple to draw conclusions about cosmological parameter estimation from, they do suggest that where correlations between  $\tau_A$ -probabilities and distance moduli exist within a class of SNe, it may be worthwhile accounting for it by using the modified posterior. Currently it is most common when modelling SNe for cosmology, to assume that the likelihood  $f_{D|\Theta,T}(d|\theta,\tau)$  is a Gaussian with unknown mean and variance,

$$D|\theta, P, T = \text{Normal}(\mu(\theta, T), \sigma(T)^2).$$

If one wishes to include the  $\tau_A$ -probabilities in the likelihood, one could include a linear shift in  $P$  for the mean or variance. That is,

$$D|\theta, P, T = \text{Normal}(\mu(\theta, T) + c_1 P, \sigma(T)^2 + c_2 P).$$

Of course this is just one possibility, and one would need to analyse SN data to get a better idea of how  $P$  should enter into the above equation.



**Figure 8.** Posterior distributions on the parameters  $(\mu_A, \mu_B)$  using the correct posterior (4) (solid) and the KBH posterior (1) (dashed). The KBH posterior assumes independence between  $D|T$  and  $P$ . Plotted are the 80%, 95% and 99% confidence levels. The true parameters (orange point) lie within the 95 % confidence regions of both posteriors.

## 7 OBTAINING UNBIASED $\tau_A$ -PROBABILITIES

In this section we investigate likely sources of  $\tau_A$ -probability biases such as those presented in Section 5, and discuss how to detect and remove them. For SNe, one source of  $\tau_A$ -probability bias could be the failure to take into account the preferential confirmation of bright objects. This type of bias has been considered in the machine learning literature under the name of selection bias, and we here present the relevant ideas from there. We end the section with a brief discussion on how one could model the pdfs  $f_{D|\Theta,P,T}$  and  $f_{D|\Theta,F,P,T}$ , which are the likelihoods appearing in the extended posteriors introduced in Section 2.

### 7.1 Selection Bias

With respect to classification methods, selection bias refers to the situation where the confirmed data is a non-representative sample of the unconfirmed data. A selection bias is sometimes also referred to as a covariate shift although the two are defined slightly differently, as described in Bickel et al. (2007). With selection bias, the confirmed data set is first randomly selected from the full set, and then at a second stage it is non-randomly reduced. Such is the situation with a population census, where at a first stage, a random sample of people is selected from the full population, and then at a second stage, people of a certain disposition cooperate more readily than others, resulting in a biased sample of respondents.

A form of selection bias which is well known in observational astronomy is the Malmquist bias, whereby magnitude limited surveys lead to the preferential detection of intrinsically bright (low apparent magnitude) objects. In the case of SN cosmology, the bias is also towards the confirming of bright SNe. A reason for this bias is that the telescope time required to accurately classify a SN is inversely proportional to the SN's brightness. It is therefore relatively cheap to confirm bright objects and expensive to confirm faint ones.

If the SN confirmation bias is ignored, certain inferences made about the global population of SNe are likely to



be inaccurate. In particular, estimates of a classifier's False Positive and False Negative Rates will be biased, and the estimated  $\tau_A$ -probabilities will be biased in certain circumstances, as we will discuss in the following section.

### 7.1.1 Formalism

Following where possible the notation of Fan et al. (2005), in what follows we assume that variables  $(X, T, F)$  are drawn from  $\mathcal{X} \times \mathcal{T} \times \mathcal{F}$ , where

- (i)  $\mathcal{X}$  is the feature space,
- (ii)  $\mathcal{T} = \{A, B\}$  is the binary type space,
- (iii)  $\mathcal{F} = \{0, 1\}$  is the binary confirmation space, where  $F = 1$  if confirmed ( $F$  for followed-up).

A realisation  $(x, \tau, f)$  lies in either the test set or the training sets, defined respectively as:

$$\begin{aligned} \text{test set} &\stackrel{\text{def}}{=} \{(x, \tau, f) \text{ s.t. } f = 0\} \\ \text{training set} &\stackrel{\text{def}}{=} \{(x, \tau, f) \text{ s.t. } f = 1\}. \end{aligned}$$

For SN cosmology it could be that  $\mathcal{X}, \mathcal{T}$  and  $\mathcal{F}$  are respectively,

- (i)  $\mathcal{X}$  is the space of all possible photometric data, where a SN's photometric data consists of apparent magnitudes and observational standard deviations in four colour bands over several nights.
- (ii)  $\mathcal{T} = \{\text{Ia}, \text{nIa}\}$ , type Ia and non-Ia SNe.
- (iii)  $\mathcal{F} = \{0, 1\}$ , where  $F = 1$  if the SN has been spectroscopically confirmed and thus has its type known.

By having a training set be unbiased we mean that it is a representative sample of the test set, specifically that  $F$  is independent of both  $X$  and  $T$ . That is, the probability of confirmation is independent of features and type:

$$P(F = 1|X = x, T = \tau) = P(F = 1). \quad (16)$$

When the training set is unbiased, training set and test set objects are drawn from the same distribution over  $\mathcal{X} \times \mathcal{T}$ . This distribution over  $\mathcal{X} \times \mathcal{T}$  can be estimated from the training set, so directly providing an estimate of the more useful test set distribution.

There are three important ways in which the independence relation (16) can break down, resulting in a biased training set, as described in Zadrozny (2004) and listed below. By removing bias from a training set, we mean reweighting the training points such that the training set becomes unbiased.

- (i) Confirmation is independent of features only when conditioned on type:  $F|T$  and  $X$  are independent. This is the simplest kind of biasing, and there are methods for correcting for it (Bishop 1996), (Elkan 2001). This is not the bias which exists in SN data.
- (ii) Confirmation is independent of type only when given features:  $F|X$  and  $T$  are independent. If the decision to confirm is based on  $X$  and perhaps some other factors which are independent of  $T$ , this is the bias which exists. This is probably the bias which exists in SN data, and there are methods for correcting for it, as we will discuss.
- (iii) Confirmation depends on both features and type simultaneously. In this case, it is not possible to remove the bias from the data unless the exact form of the bias is known.

The decision to confirm a SN can be dictated by different features, examples include Sako et al. (2008); Sullivan et al. (2006), all of which are contained in the photometric data  $X$ . Such was the also case in the SNPCC where the probability of confirmation was based entirely on the peak magnitude in the  $r$  and  $i$  f, as we will discuss in Section 8. In reality, there are other factors which affect the confirmation decision such as the weather and telescope availability, but these are independent of SN type. Therefore the type (ii) bias above is the bias which exists in the SN data. Thus, for the remainder of this section we'll assume the type (ii) bias, that is

$$P(F = 1|X = x, T = \tau) = P(F = 1|X = x). \quad (17)$$

The assumption of the type (ii) bias can be made stronger. The decision to confirm an object does not in general depend on all of  $X$  but only a low-dimensional component ( $X_F$ ) of it, and so we have

$$P(F = 1|X = x, T = \tau) = P(F = 1|X_F = x_F), \quad (18)$$

where  $X_F$  is contained in  $X$ . For SNe,  $X_F$  could be the peak apparent magnitude in certain colour bands.

In the following subsection we will describe how to correctly obtain  $\tau_A$ -probabilities under the assumption of a bias described by (18).

## 7.2 Correctly obtaining $\tau_A$ -probabilities

Let us remind the reader as to how we defined  $\tau_A$ -probabilities in the introduction:

$$\tau_A\text{-probability} \stackrel{\text{def}}{=} P(T_i = A|X_{P,i} = x_{P,i}) = p_i, \quad (19)$$

where  $X_{P,i}$  is an observable feature of the  $i$ th object, extracted from  $X_i$ . Estimates of  $p_i$  values can be obtained using several methods, of which those mentioned previously are Poznanski et al. (2002); Newling et al. (2011); Richards et al. (2011); Guy et al. (2007) It is worth rementioning that these different methods attempt to estimate different probability functions, as they each condition on different SN features. Thus there is no sense in which one set of  $\tau_A$ -probabilities estimates is *the* correct set.

We now make an adjustment to definition (19), to take into account that biased follow-up may result in an additional conditional dependence on  $F$ :

$$\tau_A\text{-probability} \stackrel{\text{def}}{=} P(T_i = A|F_i = f_i, X_{P,i} = x_{P,i}) = P_i. \quad (20)$$

The most informative  $\tau_A$ -probabilities one could use would be those conditional on all of the features at one's disposal,

$$X_P = X : \quad p_i = P(T = A|F = 0, X = x). \quad (21)$$

However, when  $\mathcal{X}$  is a high-dimensional non-homogeneous space, as is the case with photometric SN data, it can be difficult to approximate (21) accurately. It is for this reason that it is necessary to reduce the features to a lower dimensional quantity  $X_P \in \mathcal{X}_P$ , so that the  $\tau_A$ -probabilities are calculated from a subspace ( $\mathcal{X}_P$ ) of the full feature space, as described by (20). The subspace  $\mathcal{X}_P$  should be chosen to retain as much type specific information as possible while being of a sufficiently

low dimension. In the SNPCC Newling et al. (2011) chose  $\mathcal{X}_P$  to be a 20-dimensional space of parameters obtained by fitting lightcurves.

The job of obtaining estimated  $\tau_A$ -probabilities for test set objects ( $F = 0$ ) is one of obtaining an estimate of the type probability mass function,

$$f_{T|F,X_P}. \quad (22)$$

Again, for (22) we prefer not to use the standard mass function notation, in order to to neaten certain integrals which follow. The  $\tau_A$ -probability of a test set object can now be expressed in the following way,

$$P(T = A|F = 0, X_P = x_P) = f_{T|F,X_P}(A|0, x_P).$$

Using kernel density estimation, boosting, or any other method of approximating a probability function, one can construct an approximation ( $\hat{f}$ ) of the type probability function for training set objects,

$$\hat{f}(x_P) \approx f_{T|F,X_P}(A|1, x_P). \quad (23)$$

Using the estimate  $\hat{f}$  in (23) one can estimate the  $\tau_A$ -probabilities for the training set objects:

$$P(T = A|F = 1, X_P = x_P) \approx \hat{f}(x_P). \quad (24)$$

The estimate (24) is not directly important as the training set object types are known exactly. But it is only through the training set objects that we can learn anything about the types of the test set objects.

How  $\hat{f}$  from the training set is related to  $f_{T|F=0,X_P}$  (22) depends on the relationship between  $X_F$  (the data which determines confirmation probability) and  $X_P$  (the data used to calculate  $\tau_A$ -probabilities). There are two cases to consider. The first, which we write as  $\mathcal{X}_F \subset \mathcal{X}_P$ , is when the data which determines confirmation probabilities is completely contained in the data used to calculate  $\tau_A$ -probabilities. That is,

$$\mathcal{X}_F \subset \mathcal{X}_P \quad \stackrel{\text{def}}{\iff} \quad P(F = 1|X_P = x_P) = P(F = 1|X_F = x_F).$$

The second case, when  $\mathcal{X}_F \not\subset \mathcal{X}_P$  is when not all confirmation information is contained in  $X_P$ ,

$$\mathcal{X}_F \not\subset \mathcal{X}_P \quad \stackrel{\text{def}}{\iff} \quad P(F = 1|X_P = x_P) \neq P(F = 1|X_F = x_F).$$

In the case of  $\mathcal{X}_F \subset \mathcal{X}_P$ , it can be shown that,

$$P(F = 1|T = \tau, X_P = x_P) = P(F = 1|X_F = x_F). \quad (25)$$

### 7.2.1 $\mathcal{X}_F \subset \mathcal{X}_P$

We will show that in the case of  $\mathcal{X}_F \subset \mathcal{X}_P$ , a type probability function approximating the training population ( $\hat{f}$ ) is an unbiased approximation for the type probability function of the test population ( $F = 0$ ). To show this we start with the type probability of a test object:

$$P(T = \tau|F = 0, X_P = x_P).$$

→ Using Bayes' Theorem, we have

$$= \frac{P(F = 0|T = \tau, X_P = x_P) \cdot P(T = \tau|X_P = x_P)}{P(F = 0|X_P = x_P)}$$

→ Then using (25), we have

$$= \frac{P(F = 0|X_F = x_F) \cdot P(T = \tau|X_P = x_P)}{P(F = 0|X_F = x_F)} \\ = P(T = \tau|X_P = x_P). \quad (26)$$

→ Using the same steps as above but in reverse and with  $F = 1$ , we arrive at

$$= P(T = \tau|F = 1, X_P = x_P).$$

→ This is the type probability function for training set objects, and it can be approximated:

$$\approx \hat{f}(x_P). \quad (27)$$

This is a useful result, as it says that  $\hat{f}$  is not only an approximation of the type probability function of the training data, but also of the test set. Thus,  $\hat{f}$  should provide unbiased  $\tau_A$ -probabilities for the test set when  $\mathcal{X}_F \subset \mathcal{X}_P$ .

It should be noted that for  $\hat{f}$  to be a good approximation for the test set, it is necessary that the training set covers all regions of  $\mathcal{X}_P$  where there are test points. That is, if there are values of  $x_P$  for which  $P(X_P = x_P|F = 1) = 0$  and  $P(X_P = x_P|F = 0) \neq 0$ , then the approximation  $\hat{f}$  will not converge to  $f_{T|F=0,X_P}$  as the training set size grows. One can refer to Fan et al. (2005) for a full treatment of this topic.

With respect to SNe, the requirement of the preceding paragraph is that, if a SN is too faint to be confirmed and to enter the training set, it should not enter the test set either. We will return to this point again in Section 8.

One important question which we do not attempt to answer here is, how many SNe of different apparent magnitudes should be confirmed to obtain as rapid as possible convergence of  $\hat{f}$  to  $f_{T|F=0,X_P}$ . An interesting method for deciding which SNe to confirm may be one based on the real-time approach proposed in Freund et al. (1997), where the decision to add an object to the training set is based on the uncertainty of its type using the currently fitted  $\hat{f}$ . In Section 8 we discuss this further.

### 7.2.2 $\mathcal{X}_F \not\subset \mathcal{X}_P$

If  $\mathcal{X}_F \not\subset \mathcal{X}_P$  we will not be able to use  $\hat{f}$  to estimate the  $\tau_A$ -probabilities in the test set, as (27) required  $\mathcal{X}_F \subset \mathcal{X}_P$ . In addition to this problem of not being able to use  $\hat{f}$  to obtain unbiased  $\tau_A$ -probabilities for the test set objects, if  $\mathcal{X}_F \not\subset \mathcal{X}_P$  then

$$P(T = \tau|X_P = x_P) \neq P(T|X_P = x_P, X_F = x_F).$$

This tells us that there is additional type information to be obtained from  $X_F$ , and so by not including  $X_F$  one is wasting type information. For this reason we recommend reconstructing the  $\tau_A$ -probabilities based on redefined features,  $X_P \leftarrow (X_F, X_P)$ .

However, it is possible that one explicitly does not want to use  $X_F$  in calculating  $\tau_A$ -probabilities. This may be the case if one wishes to reduce the dependence between  $D$  and  $P$ , as presented in Section 2. For SNe, this may involve obtaining  $\tau_A$ -probabilities from shape alone, independent of magnitude, so that  $\mathcal{X}_P$  is a space whose dimensions describe only shape and not magnitude. In this case, as we cannot

use  $\hat{f}$ , we need to use the relationship derived in Shimodaira (2000),

$$\begin{aligned} P(T = \tau | F = 0, X_P) \\ = \int_{\mathcal{X}_F} f_{T, X_F | F, X_P}(\tau, x_F | 1, x_P) \cdot w(x_F, x_P) dx_F, \end{aligned} \quad (28)$$

where the weight function is defined as

$$w(x_F, x_P) = \frac{f_{F|X_F}(0|x_F)f_{F|X_P}(1|x_P)}{f_{F|X_F}(1|x_F)f_{F|X_P}(0|x_P)}. \quad (29)$$

Notice that if  $\mathcal{X}_F \subset \mathcal{X}_P$ , then  $w(x_F, x_P) = 1$  and so (28) reduces to the type probability function for training set objects, approximated by  $\hat{f}$  as expected from (27). When  $w(x_F, x_P) \neq 1$ , the training set type probability function  $\hat{f}$  cannot be used directly as an approximation to the test set type probability function. However, if each training set object is weighted using (28), then an unbiased test set type probability function approximation can be obtained.

The weight function (29) does not require any type information and so can be estimated as a first step. This additional step of estimation introduces additional error into the final estimate of (22), a theoretical analysis of which is presented in Cortes et al. (2008). An alternative to the two-stage approach would be to fit the two terms in (28) simultaneously, as suggested and described by (Bickel et al. 2007). The use of (29) was first suggested in Shimodaira (2000), where a detailed analysis of the asymptotic behaviour of its approximation is given. Therein, it is suggested that (29) be approximated by kernel density estimation.

In the case where  $F$  and  $X_P$  are independent, the weight function reduces to one of only  $X_F$ ,

$$w(X_F = x_F) = \frac{P(F = 0 | X_F = x_F)P(F = 1)}{P(F = 1 | X_F = x_F)P(F = 0)}. \quad (30)$$

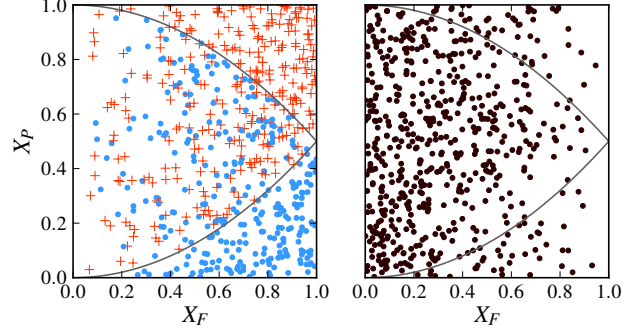
This reduction in dimension may be valuable in approximating the weight function.

### 7.3 Detecting and removing biases in $\tau_A$ -probabilities

In the previous section we presented the correct way in which to estimate  $\tau_A$ -probabilities in the case  $\mathcal{X}_F \not\subset \mathcal{X}_P$ . In this section we will present an example illustrating this process, but in the context of bias removal.

Suppose that we have a program which outputs scalar values ( $\tilde{p}$ ), which are purported  $\tau_A$ -probabilities. We believe that the output values have some unspecified bias, which we wish to remove. An assumption we make is that the  $\tilde{p}$  values are calculated in the same way for training and test sets. That is that the program does not process cases  $F = 0$  and  $F = 1$  differently. It may seem strange to be interested in what the program does when  $F = 1$ , but as already mentioned it is only from the training set that we can learn anything about the test set. The idea now is to treat the received  $\tilde{p}$  values as the  $x_P$ s from the previous section, and not directly as  $\tau_A$ -probabilities.

For this example, we choose  $\mathcal{X}_F = [0, 1]$ . To now transform a test set value  $\tilde{p} \in [0, 1]$  into an unbiased  $\tau_A$ -probability using (28), one needs to estimate certain probability functions using kernel density estimation.



**Figure 9.** Realisations of a training set (left) containing type A (red pluses) and type B (blue points) objects, and a test set (right), drawn according to (31). Overlaid are faint lines delineating the discrete regions described by (31)

tion. The necessary functions we see from (28) and (29) are  $f_{T, X_F | F, X_P}(\tau, x_F | 1, \tilde{p})$ ,  $f_{F|X_F}(1, x_F)$ ,  $f_{F|X_P}(0, \tilde{p})$ ,  $f_{F|X_F}(0, x_F)$  and  $f_{F|X_P}(0, x_P)$ .

It is an interesting and important question as to how accurately these probability functions can be approximated with few data points, but for this example we assume them known,

$$\begin{aligned} f_{T, X_F | F, X_P}(A, x_F | 1, \tilde{p}) &= \begin{cases} x_F & \text{if } \frac{1}{2}x_F^2 < \tilde{p} < 1 - \frac{1}{2}x_F^2, \\ 2 \cdot x_F & \text{if } \tilde{p} > 1 - \frac{1}{2}x_F^2 \\ 0 & \text{if } \tilde{p} < \frac{1}{2}x_F^2. \end{cases} \\ f_{F|X_F}(0, x_F) &= (1 - x_F), \\ f_{F|X_F}(1, x_F) &= x_F, \\ f_{F|X_P}(1|\tilde{p}) &= f_{F|X_P}(0|\tilde{p}) = \frac{1}{2}. \end{aligned} \quad (31)$$

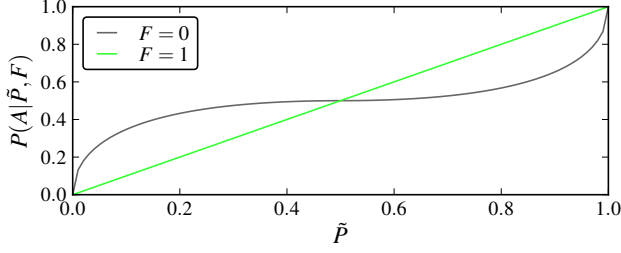
Realisations from the above distribution are illustrated in Figure 9. By integrating  $x_F$  out of  $f_{T, X_F | F, X_P}(A, x_F | 1, \tilde{p})$  in (31), we have that

$$P(T = A | F = 1, X_P = \tilde{p}) = \tilde{p}. \quad (32)$$

That is, in the training set  $\tilde{p}$  is an unbiased estimate of a  $\tau_A$ -probability. The  $\tau_A$ -probabilities for objects in the test set we estimate using (28),

$$\begin{aligned} P(T = A | F = 0, X_P = \tilde{p}) \\ = \int_{\mathcal{X}_F} f_{T, X_F | F, X_P}(\tau, x_F | 0, x_P) dx_F \\ = \int_{\mathcal{X}_F} f_{T, X_F | F, X_P}(\tau, x_F | 1, x_P) w(x_F, \tilde{p}) dx_F \\ = \int_{\mathcal{X}_F} f_{T, X_F | F, X_P}(\tau, x_F | 1, x_P) \frac{1 - x_F}{x_F} dx_F \\ = \begin{cases} \sqrt{2\tilde{p}} - \tilde{p} & \text{if } \tilde{p} < 0.5, \\ 2 - \tilde{p} - \sqrt{2 - 2\tilde{p}} & \text{if } 0.5 < \tilde{p}. \end{cases} \end{aligned} \quad (33)$$

The  $\tau_A$ -probabilities (32) and (33) are plotted in Figure 10, where we see that  $\tilde{p}$  provided accurate  $\tau_A$ -probabilities for the training set, but not for the test set. This is not unexpected in reality, where the program providing the  $\tau_A$ -probabilities may have been trained only on the biased training data. It is important to remember that this bias should only arise when  $\mathcal{X}_F \not\subset \mathcal{X}_P$ .



**Figure 10.** Corrected  $\tau_A$ -probabilities. The disproportionately large number of training SNe with decisive  $\tau_A$ -probabilities (as depicted in Figure 9), causes  $\tilde{p}$  values to be too confident as test set  $\tau_A$ -probability estimates.

## 8 SUPERNOVA SURVEYS AND THE SNPCC

The SNPCC provided a simulated spectroscopic training data set of approximately 1000 known SNe. The challenge was then to predict the types of approximately 20 000 other objects<sup>3</sup> from their lightcurves alone. Since the end of the competition, the types of all the simulated SNe have been released, making a post competition autopsy relatively easy to perform. In the results paper Kessler et al. (2010) we see that the probability that a SN was confirmed was based on the  $r$ -band and  $i$ -band quantities,

$$\epsilon_{\text{spec}}^{\text{band}} = \epsilon_0 (1 - x^i) \quad x \stackrel{\text{def}}{=} \frac{m_{\text{peak}}^{\text{band}} - M_{\text{min}}^{\text{band}}}{m_{\text{lim}}^{\text{band}} - M_{\text{min}}^{\text{band}}},$$

where  $m_{\text{peak}}^{\text{band}}$  is the band-specific apparent magnitude of a SN, and  $M_{\text{min}}^{\text{band}}$  and  $m_{\text{lim}}^{\text{band}}$  are constants. In Kessler et al. (2010) it is given that for  $\text{col} = r$  and  $\text{col} = i$ ,

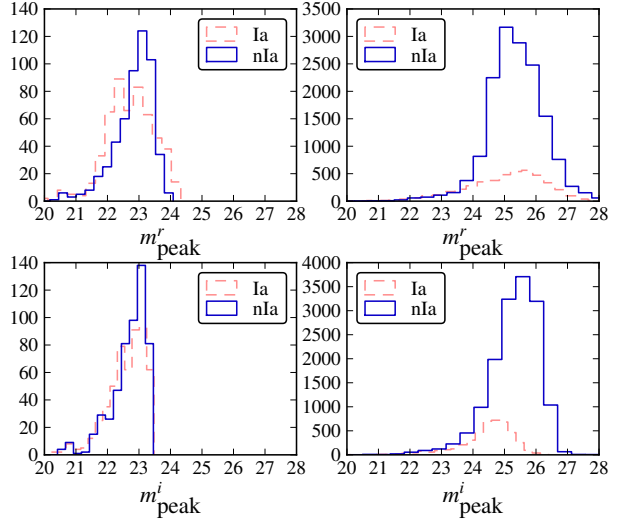
$$\begin{aligned} \epsilon_{\text{spec}}^r &= \epsilon_0 (1 - x^5) & x &\stackrel{\text{def}}{=} \frac{m_{\text{peak}}^r - 16.0}{5.5} \\ \epsilon_{\text{spec}}^i &= \epsilon_0 (1 - x^6) & x &\stackrel{\text{def}}{=} \frac{m_{\text{peak}}^i - 21.5}{2.0} \end{aligned} \quad (34)$$

where  $\epsilon_0$  is some constant. Once  $\epsilon_{\text{spec}}^i$  and  $\epsilon_{\text{spec}}^r$  have been calculated, if a  $[0 \rightarrow 1]$  uniform random number is less than either of them, confirmation is performed. As confirmation depends only on  $\epsilon_{\text{spec}}^i$  and  $\epsilon_{\text{spec}}^r$ , we have from (26) that

$$P(T = \tau | F = 0, m_{\text{peak}}^i, m_{\text{peak}}^r) = P(T = \tau | F = 1, m_{\text{peak}}^i, m_{\text{peak}}^r). \quad (35)$$

Equation (35) can be interpreted as saying that the ratio Ia:nIa is the same in a given  $m_{\text{peak}}^i, m_{\text{peak}}^r$  bin. The manner in which the follow-up was simulated should of course guarantee that (35) holds. In theory one should be able to deduce the verity of (35) from Figure 11, but the redshift bins with large numbers of confirmed SNe are too sparsely populated by unconfirmed SNe to check that the Ia:nIa is invariant. To be in a position where (35) can be checked is in general an unrealistic luxury, as without the types of the test objects this is impossible.

In terms of obtaining accurate  $\tau_A$ -probabilities, a disturbing feature of Figure 11 is the absence of training SNe with high apparent magnitudes. With no training SNe with



**Figure 11.** Counts of confirmed (left) and not confirmed (right) SNe, Ia (dashed) and non-Ia (solid) as a function of  $m_{\text{peak}}^r$  (above) and  $m_{\text{peak}}^i$  (below).

$i$ -band apparent magnitudes greater than 23.5, we cannot infer the types of test SNe with apparent magnitudes greater than 23.5. Indeed there would be no non-astrophysical reason not to believe that all SNe with apparent magnitudes greater than 23.5 are non-Ia. As already mentioned in Section 7.1, in situations where the training set does not span the test set, one should ignore unrepresented test objects from all analyses. All test SNe other than those for which there are training SNe of comparable peak apparent magnitudes in  $r$  and  $i$  bands should be removed from a BEAMS analysis, unless there is a valid astrophysical reason not to do so. This entails ignoring about 95% of unconfirmed SNe; an enormous cut. We therefore consider it important to confirm more faint SNe.

In Newling et al. (2011), a comparison is made between training a boosting algorithm on the non-representative spectroscopically confirmed SNe and a representative sample, randomly selected from the unconfirmed SN set. Therein, the authors use twenty fitted lightcurve parameters, including fitted apparent magnitudes in  $r$  and  $i$  bands. This corresponds to the situation discussed in Section 7.2.1, where  $\mathcal{X}_P \subset \mathcal{X}_F$ . For this reason, the probability density function  $\hat{f}$  in 24 as estimated by their boosting algorithm should be an unbiased estimate for  $f_{T|F=0, \mathcal{X}_P}$ . But being unbiased does not guarantee low error, and when trained on the confirmed SNe, regions of parameter space corresponding to high apparent magnitude had no training SNe with which to learn, and so the approximation of 22 was poor. However when trained on the representative set, every region of populated parameter space was represented by the training set, and the approximation of 22 was greatly improved.

In their paper, Richards et al. (2011) describe their entry in the SNPCC, and they report how a semi-supervised learning algorithm performs better with a few faint training SNe than with many bright ones. The comparison was per-

<sup>3</sup> These lightcurves are available at [http://sdssdp62.fnal.gov/sdssn/SIMGEN\\_PUBLIC/](http://sdssdp62.fnal.gov/sdssn/SIMGEN_PUBLIC/)

formed while keeping the total confirmation time constant. Thus their conclusion was the same as ours; that it is important to obtain a more representative SN training sample.

## 9 CONCLUSIONS AND RECOMMENDATIONS

In this paper we discussed BEAMS, and extended the KBH posterior probability function to the case when  $D|T$  (distance modulus | type) and  $P$  (type probability) are dependent. In Section 6 we considered an example where the dependence between  $D|T$  and  $P$  is strong, and observed a large reduction in the posterior width using the extended posterior as opposed to the KBH posterior. No bias is observed when using either the extended or the KBH posterior.

In Section 4 we considered examples where the KBH posterior is valid, that is when  $D|T$  and  $P$  are independent. We performed tests to ascertain the importance to BEAMS of i) the decisiveness of the  $\tau_A$ -probabilities (observations of  $P$ ), and ii) sample size. In one test (4.1), we observed how doubling a sample size reduces error in parameter estimation more than obtaining the true type identity of the objects does. In another test (4.2), we observed how BEAMS accurately locates two population means, but fails to match each mean to its population.

We looked at the effects of using biased  $\tau_A$ -probabilities in Section (5). The result of KBH, that  $\tau_A$ -probability biases towards population  $A$  affect the population's parameter estimates less than biases in favour of population  $B$ , was observed. A similar result which is uncovered is that biases towards high decisiveness are more damaging than biases towards low decisiveness. In other words, it is better to be conservative in your prior type beliefs than too confident.

Our recommendations for BEAMS may thus be summarised as follows. Firstly, the inclusion in the likelihood function of  $\tau_A$ -probabilities can dramatically reduce the width of the final posterior, providing tighter constraints on cosmological parameters. Secondly, conservative estimation of  $\tau_A$ -probabilities is less harmful than too decisive an estimation. Thirdly, it is possible to remove biases in  $\tau_A$ -probabilities using the techniques described in Section (7).

In Section 7 we considered the problem of debiasing  $\tau_A$ -probabilities. Interpreting recent results from the machine learning literature in terms of SN cosmology, we discussed the different ways in which training sets can be biased and how to remove such biases. The key to understanding and correcting biases is the relationship between  $\mathcal{X}_F$  and  $\mathcal{X}_P$ , where  $\mathcal{X}_F$  are object features which determine confirmation probability, and  $\mathcal{X}_P$  are those features which determine  $\tau_A$ -probabilities. In brief, when  $\mathcal{X}_P$  contains  $\mathcal{X}_F$ ,  $\tau_A$ -probabilities should be unbiased, but if this is not the case, there are sometimes ways for correcting the bias.

With respect to future SN surveys, we emphasize the importance of an accurate record as to what information is used when deciding whether or not a SN is confirmed. Using this information, one should in theory be able to remove all the affects of selection bias when  $\mathcal{X}_F \not\subset \mathcal{X}_P$ . In other words, using all the variables which are considered in deciding whether to follow-up a SN, it will always be possible to obtain unbiased  $\tau_A$ -probabilities, irrespective of what the  $\tau_A$ -probabilities are based on. Such follow-up variables may

include early segments of light curves,  $\chi^2$  goodness of fits, fit probabilities, host galaxy position and type, expected peak apparent magnitude in certain filters, etc.

Our second recommendation for SN surveys is that more faint objects are confirmed. While it not necessary for most machine learning algorithms to have a spectroscopic training set which is exactly representative of the photometric test set, it is necessary that the spectroscopic set at least covers the photometric set. Thus having large numbers of faint unconfirmed objects without any confirmed faint objects is suboptimal.

## 10 ACKNOWLEDGEMENTS

JN has a SKA bursary and MS is funded by a SKA fellowship. BB acknowledges funding from the NRF and Royal Society. MK acknowledges financial support by the Swiss NSF. RH acknowledges funding from the Rhodes Trust.

## APPENDIX A: POSTERIOR TYPE PROBABILITIES

We here derive the posterior type probabilities based on the modifications of Section 2. The posterior type probability will be derived, conditional on  $\mathbf{D}$  and  $\mathbf{P}$ . This derivation can be easily extended to posterior type probabilities conditional on  $\mathbf{D}, \mathbf{F}$  and  $\mathbf{P}$ .

$$\begin{aligned} f_{T_i|\mathbf{D},\mathbf{P}}(A|\mathbf{d},\mathbf{p}) &= \int_{\theta} f_{T_i|\Theta,\mathbf{D},\mathbf{P}}(A|\theta,\mathbf{d},\mathbf{p}) f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p}) d\theta \\ &= \int_{\theta} f_{T_i|\Theta,D_i,P_i}(A|\theta,d_i,p_i) f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p}) d\theta \end{aligned}$$

→ we have assumed that the objects are independent,

$$= \int_{\theta} \frac{f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,A) f_{T_i|\Theta,P_i}(A|\theta,p_i)}{f_{D_i|\Theta,P_i}(d_i|\theta,p_i)} \times f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p}) d\theta$$

→ we have used Bayes' Theorem,

$$= \int_{\theta} \left( \frac{A_i}{A_i + B_i} \right) f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p}) d\theta. \quad (\text{A1})$$

→ where  $A_i = P(d_i|\theta,p_i,T_i = A)p_i$ ,  $B_i = P(d_i|\theta,p_i,T_i = B)(1 - p_i)$ , and we have assumed used that  $f_{T_i|\Theta,P_i}(A|\theta,p_i) = p_i$ .

If the posterior  $f_{\Theta|\mathbf{D},\mathbf{P}}$  confines  $\theta$  to a region sufficiently small such that  $A_i$  and  $B_i$  are approximately constant, then the posterior type probability (A1) is well approximated by  $A_i(\hat{\theta}) / (A_i(\hat{\theta}) + B_i(\hat{\theta}))$  where  $\hat{\theta}$  is the maximum likelihood estimator of  $f_{\Theta|\mathbf{D},\mathbf{P}}(\theta|\mathbf{d},\mathbf{p})$ . Furthermore, the posterior odds ratio,

$$\text{posterior odds ratio} \stackrel{\text{def}}{=} \frac{f_{T_i|\mathbf{D},\mathbf{P}}(A|\mathbf{d},\mathbf{p})}{f_{T_i|\mathbf{D},\mathbf{P}}(B|\mathbf{d},\mathbf{p})}$$

can be shown to be given by the prior odds ratio multiplied by the Bayes Factor,

$$\text{posterior odds ratio} = \left( \frac{p_i}{1 - p_i} \right) \times \left( \frac{f_{D_i|\Theta, P_i, T_i}(d_i|\hat{\theta}, p_i, A)}{f_{D_i|\Theta, P_i, T_i}(d_i|\hat{\theta}, p_i, B)} \right).$$

## APPENDIX B: ADDITIONAL CONDITIONING ON THE CONFIRMATION OF SUPERNOVA TYPE

In this paper we did not distinguish between the contributions of unconfirmed and confirmed objects to the posterior. While we can calculate approximate  $\tau_A$ -probabilities for confirmed objects, these values should not enter the posterior, but be replaced by 0 (if type B) or 1 (if type A). Let us introduce the random variable  $F$  to denote whether an object is confirmed, so that  $F = 1$  if confirmed and  $F = 0$  if unconfirmed. With this introduced, we wish to replace the  $\tau_A$ -probabilities  $\mathbf{p}$  by  $\bar{\mathbf{p}}$ , where,

$$\bar{p}_i = \begin{cases} p_i & \text{if } f_i = 0, \\ 1 & \text{if } f_i = 1 \text{ and } \tau_i = A, \\ 0 & \text{if } f_i = 1 \text{ and } \tau_i = B. \end{cases}$$

We must be careful to let the new information which we introduce in  $\bar{\mathbf{p}}$  be absorbed elsewhere in the posterior. To this end, as we did in Section ?? we start afresh the posterior derivation, explicitly including the vector ( $\mathbf{f}$ ) which describes which objects have been followed-up. Doing this, we arrive at the following posterior distribution

$$f_{\Theta|D, F, P}(\theta|\mathbf{d}, \mathbf{f}, \mathbf{p}) \propto f_{\Theta}(\theta) \times \quad (B1)$$

$$\sum_{\tau} f_{D|\Theta, F, P, T}(\mathbf{d}|\theta, \mathbf{f}, \mathbf{p}, \tau) \prod_{\tau_i=A} \bar{p}_i \prod_{\tau_j=B} (1 - \bar{p}_j).$$

The new information ( $\mathbf{f}$ ) has been absorbed into the likelihood,  $f_{D|\dots}$ . For a particular application, one may now ask if the addition of  $\mathbf{F}$  in  $f_{D|\dots}$  is necessary. We have already mentioned that for SNe  $\mathbf{D}|\theta, \mathbf{T}$  is unlikely to be independent of  $\mathbf{P}$ . It is also unlikely that  $\mathbf{D}|\theta, \mathbf{T}$  is independent of  $\mathbf{F}$ , as bright SNe, which have lower fitted distance moduli at a given redshift, are confirmed more regularly than faint ones. However, it is possible that by additionally conditioning  $\mathbf{D}$  on  $\mathbf{P}$  this confirmation dependence is broken, so that  $\mathbf{D}|\theta, \mathbf{P}, \mathbf{T}$  and  $\mathbf{F}$  are independent. We leave this as an open question.

In the case of independent SNe, the posterior (B1) reduces to

$$f_{\Theta|D, F, P}(\theta|\mathbf{d}, \mathbf{f}, \mathbf{p}) \propto \prod_{i=1}^N [f_{D_i|\Theta, F_i, P_i, T_i}(d_i|\theta, f_i, p_i, A) \bar{p}_i + f_{D_i|\Theta, F_i, P_i, T_i}(d_i|\theta, f_i, p_i, B) (1 - \bar{p}_i)]. \quad (B2)$$

## REFERENCES

Abazajian K. N., Adelman-McCarthy J. K., Agüeros M. A., Allam S. S., Allende Prieto C., An D., Anderson K. S. J., Anderson S. F., Annis J., Bahcall N. A., et al. 2009, ApJS, 182, 543

Bickel S., Brückner M., Scheffer T., 2007, in ICML '07: Proceedings of the 24th international conference on Machine learning Discriminative learning for differing training and test distributions. ACM, New York, NY, USA, pp 81–88  
 Bishop C. M., 1996, Neural Networks for Pattern Recognition, 1st edn. Oxford University Press, USA  
 Cortes C., Mohri M., Riley M., Rostamizadeh A., 2008, CoRR, abs/0805.2775  
 Eisenstein D. J., et al., 2005, ApJ, 633, 560  
 Elkan C., 2001, in Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence The foundations of cost-sensitive learning. pp 973–978  
 Fan W., Davidson I., Zadrozny B., Yu P. S., 2005, in Proceedings of the Fifth IEEE International Conference on Data Mining, An improved categorization of classifiers sensitivity on sample selection bias  
 Freund Y., Seung H. S., Shamir E., Tishby N., 1997, in Machine Learning Selective sampling using the Query by Committee algorithm. pp 133–168  
 Fu L., et al., 2008, *â*, 479, 9  
 Giannantonio T., Scranton R., Crittenden R. G., Nichol R. C., Boughn S. P., Myers A. D., Richards G. T., 2008, Phys. Rev. D, 77, 123520  
 Guy J., Astier P., Baumont S., Hardin D., 2007, A&A, 466, 11  
 Johnson B. D., Crofts A. P. S., 2006, AJ, 132, 756  
 Kessler R., Conley A., Jha S., Kuhlmann S., 2010, arXiv:1001.5210  
 Kessler R., et al., 2010, PASP, 122, 1415  
 Komatsu E., et al., 2011, ApJS, 192, 18  
 Kunz M., Bassett B. A., Hlozek R. A., 2007, Phys. Rev. D, 75, 103508  
 Kuznetsova N. V., Connolly B. M., 2007, ApJ, 659, 530  
 Mantz A., Allen S. W., Rapetti D., Ebeling H., 2010, MNRAS, 406, 1759  
 Newling J., Varughese M., Bassett B., Campbell H., Hlozek R., Kunz M., Lampeitl H., Martin B., Nichol R., Parkinson D., Smith M., 2011, MNRAS, 414, 1987  
 Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, MNRAS, 381, 1053  
 Percival W. J., et al., 2010, MNRAS, 401, 2148  
 Perlmutter S., et al., 1999, ApJ, 517, 565  
 Poznanski D., Gal-Yam A., Maoz D., Filippenko A. V., Leonard D. C., Matheson T., 2002, PASP, 114, 833  
 Poznanski D., Maoz D., Gal-Yam A., 2007, AJ, 134, 1285  
 Richards J. W., Homrighausen D., Freeman P. E., Schafer C. M., Poznanski D., 2011, ArXiv 1103.6034  
 Riess A. G., et al., 1998, AJ, 116, 1009  
 Rodney S. A., Tonry J. L., 2009, ApJ, 707, 1064  
 Sako M., Bassett B., Connolly B., Dilday B., Campbell H., Frieman J., Gladney L., Kessler R., Lampeitl H., Mariner J., Miquel R., Nichol R., Schneider D., Smith M., Sollerman J., 2011, ArXiv e-prints  
 Sako M., et al., 2008, AJ, 135, 348  
 Shimodaira H., 2000, Journal of Statistical Planning and Inference, 90, 227  
 Sullivan M., et al., 2006, AJ, 131, 960  
 Zadrozny B., 2004, in Proceedings of the twenty-first international conference on Machine learning ICML '04, Learning and evaluating classifiers under sample selection bias. ACM, New York, NY, USA, pp 114–